Jurnal Ilmu Komputer dan Informasi (Journal of a Science and Information). 11/2 (2018), 67-76
DOI: http://dx.doi.org/10.21609/jiki.v11i2.536

BAYESIAN BERNOULLI MIXTURE REGRESSION MODEL FOR BIDIKMISI SCHOLARSHIP CLASSIFICATION

Nur Iriawan[1], Kartika Fithriasari[1], Brodjol Sutija Suprih Ulama[2], Wahyuni Suryaningtyas[3], Irwan Susanto[3], and Anindya Apriliyanti Pravitasari[3]

Department of Statistics, Faculty of Mathematics, Computing, and Data Sciences, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia.
[2]Department of Business Statistics, Faculty of Vocational Studies, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia.
[3]Doctoral Candidate at Department of Statistics, Faculty of Mathematics, Computing, and Data Sciences, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia.

E-mail: nur_i@statistika.its.ac.id

Abstract

Bidikmisi scholarship grantees are determined based on criteria related to the socioeconomic conditions of the parent of the scholarship grantee. Decision process of Bidikmisi acceptance is not easy to do, since there are sufficient data of prospective applicants and variables of varied criteria. Based on these problems, a new approach is proposed to determine Bidikmisi grantees by using the Bayesian Bernoulli mixture regression model. The modeling procedure is performed by compiling the accepted and unaccepted cluster of applicants which are estimated for each cluster by the Bernoulli mixture regression model. The model parameter estimation process is done by building an algorithm based on Bayesian Markov Chain Monte Carlo (MCMC) method. The accuracy of acceptance process through Bayesian Bernoulli mixture regression model is measured by determining acceptance classification percentage of model which is compared with acceptance classification percentage of the dummy regression model and the polytomous regression model. The comparative results show that Bayesian Bernoulli mixture regression model approach gives higher percentage of acceptance classification accuracy than dummy regression model and polytomous regression model.

Keywords: *Bernoulli mixture regression model, Bayesian MCMC, Gibbs Sampling, Bidikmisi.*

Abstrak

Penerima beasiswa Bidikmisi ditentukan berdasarkan kriteria yang berkaitan dengan kondisi sosial ekonomi dari orang tua atau wali dari penerima beasiswa. Proses penentuan penerimaan tidak mudah dilakukan mengingat terdapat data calon pendaftar yang cukup besar dan variabel kriteria yang bervariasi. Berdasarkan permasalahan tersebut, diusulkan pendekatan baru untuk penentuan penerima Bidikmisi dengan menggunakan model regresi mixture Bernouuli Bayesian. Prosedur pemodelan dilakukan dengan menyusun klaster pendaftar yang diterima dan tidak diterima yang selanjutnya dilakukan estimasi model untuk setiap klaster melalui model mixture regresi Bernoulli. Proses estimasi parameter model dilakukan dengan menyusun suatu algoritma dengan berdasarkan metode Bayesian Markov Chain Monte Carlo (MCMC). Ketepatan proses klasifikasi melalui model regresi mixture Bernoulli Bayesian diukur dengan menentukan prosentase klasifikasi penerimaan dari model yang dibandingkan dengan prosentase klasifikasi penerimaan dari model regresi dummy dan model regresi polytomous. Hasil perbandingan menunjukkan bahwa pendekatan model regresi nixture Bernoulli Bayesian memberikan prosentase ketepatan klasifikasi penerimaan lebih tinggi dibanding model regresi dummy dan model regresi polytomous.

Kata Kunci: *Model mixture regresi Bernoulli, Bayesian MCMC, Gibbs Sampling, Bidikmisi.*

1. Introduction

In the era of Asean Economy Community (AEC) in 2015, education occupies the front guard in the development of Human Resources. AEC enforcement becomes a momentum to make improvements in Indonesia's education sector to be able to produce human resources that have high competitiveness. The Government in order to improve people's productivity and competitive-ness in the international market launched Bidik-misi Education Cost Assistance

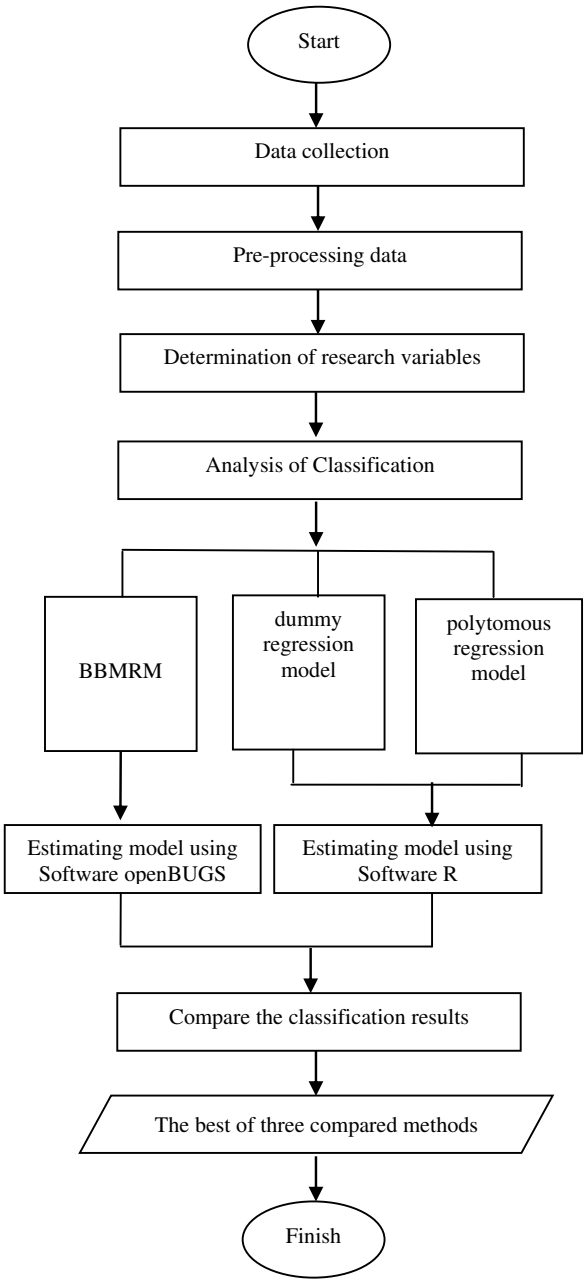

Figure 1. Flowchart Classification Bidikmisi using BBMRM, Dummy Regression Model, and Polytomous Regression Model.

Program [1]. Bidikmisi implementation is given especially for marginal people. However, there are indications of problems in the implementation of the program on the acceptance of Bidikmisi scholarships for Higher Education, i.e. the existence of unaccept-able acceptance conditions.

Data response status acceptance Bidikmisi is in binary type (0 and 1), so by involving the founder covariate Bidikmisi scholarship recipient that is the main criteria factor Parent Revenue and Total Household Counts House produces data distributed Bernoulli mixture two components. The characteristics of each component of the Bernoulli mixture can be identified through the Bernoulli Mixture modeling by involving the Bidikmisi scholarship finder covariates. The mixture

TABLE 1
IDENTIFICATION COMPONENTS MIXTURE OF BIDIKMISI SCHOLARSHIP 2015

Y	CFC	AC	Condition	Interpretation
1	0	0	wrong	Acceptance condition is wrong (AC = 0) when the applicants who receive scholarships (Y = 1) have the category of wealthy family (CFC = 0).
0	1	0	wrong	Acceptance condition is wrong (AC = 0) when the applicants who do not receive scholarships (Y = 0) have the category of poor family (CFC = 1).
1	1	1	right	Acceptance condition is right (AC = 1) when the applicants who receive scholarships (Y = 1) have the category of poor family (CFC = 1).
0	0	1	right	Acceptance condition is right (AC = 1) when the applicants who do not receive scholarships (Y = 0) have the category of wealthy family (CFC = 0).

component obtained corresponds to the acceptance condition of Bidikmisi. The condition is identified as the category of acceptance with the correct conditions and the acceptance with the wrong conditions. Acceptance is categorized as true if the students are accepted Bidikmisi with conditions of incapacity and vice versa. Furthermore, the acceptance is categorized incorrectly if the student is accepted Bidikmisi with capable conditions and vice versa. Characteristics of each component can be identified through the Bernoulli Mixture Model (BMM). Therefore, it is necessary to analyze the classification of Bidikmisi acceptance based on the mixture component.

This study aims to perform the analysis of Bidikmisi acceptance classification. The proposed Bernoulli Mixture regression method couple with Bayesian MCMC approach for classification, which is run in openBUGS software, would be compared with dummy regression and polytomous regression. The dummy regression analysis classifies Bidikmisi acceptance based on the criteria of being capable and incapable. Meanwhile, polytomous regression analysis classifies Bidikmisi acceptance using four acceptance criteria. These last two approaches are run in software R.

Gibbs Sampler Algorithm :

Given $\boldsymbol{\Theta} = (\Theta_1, \Theta_2, ..., \Theta_d)' = (\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_L, \boldsymbol{\pi})'$

1. Set initial values $\boldsymbol{\Theta}^{(0)}$.
2. For $t = 1, 2, ..., T$ repeat the following steps
   a. $\boldsymbol{\Theta} = \boldsymbol{\Theta}^{(t-1)}$
   b. For $j = 1, ..., d$ update Θ_j from
      $\Theta_j \sim f(\Theta_j \mid \boldsymbol{\Theta}_{\setminus j}, \mathbf{y})$
      $f(\Theta_j \mid \boldsymbol{\Theta}_{\setminus j}, \mathbf{y})$ is a full conditional posterior distribution and
      $\boldsymbol{\Theta}_{\setminus j} = (\Theta_1, ..., \Theta_{j-1}, \Theta_{j+1}, ..., \Theta_d)^T$
   c. $\boldsymbol{\Theta}^{(t)} = \boldsymbol{\Theta}$ save it as the generated set of values at t + 1 iteration.

Figure 2. Gibbs sampler for general BMRM.

This paper shows the performance of Bernoulli mixture regression model couple with Bayesian MCMC approach to represent the acceptance condition of Bidikmisi scholarship in provinces of Indonesia. Therefore, the research focuses on the development of Bidikmisi acceptance model using Bayesian Bernoulli Mixture Regression Model (BBMRM).

Bernoulli Mixture Model

Bernoulli Mixture Model (BMM) is frequently used in text mining [2]. BMM was firstly performed by Duda and Hart [3]. In their development, they applied BMM to various aspects of life i.e. on the study recognition of image, clustering of text and word [4, 5, 6, 7, 8, 9], on cancer and schizophrenia [10, 11, 12, 13] and in the machine learning research [14, 15].

If $Y_1, Y_2, \ldots, Y_n$ is a random sample of size n, where Y_i is a D dimension random vector with the density function of the mass of $p(y_i)$ on $\mathbf{R}^D$. Thus, it practically contains random variables corresponding to D measurements made on the i observation of some features of the phenomenon

BBMRM Algorithm:

Given $\boldsymbol{\Theta} = (\Theta_1, \Theta_2, ..., \Theta_d)'$

$= (\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_L, \boldsymbol{\pi})'$

1. Set initial values $\boldsymbol{\Theta}^{(0)}$.
2. For $t = 1, 2, ..., T$ repeat the following steps
   a. $\boldsymbol{\Theta} = \boldsymbol{\Theta}^{(t-1)}$
   b. For $j = 1, ..., d$ update β_j and π_j from
      $\beta_j \sim N(\mu, \sigma)$
      $\pi_j \sim Dirichlet(\vartheta)$
      Calculate $\text{logit}_1[g_1(\mathbf{x})]$
      Calculate $\text{logit}_2[g_2(\mathbf{x})]$

$\boldsymbol{\Theta}_{\setminus j} = (\Theta_1, ..., \Theta_{j-1}, \Theta_{j+1}, ..., \Theta_d)^T$

   c. $\boldsymbol{\Theta}^{(t)} = \boldsymbol{\Theta}$ save it as the generated set of values at t + 1 iteration.

Figure 3. Gibbs sampler for BBMRM with normal prior on β_j and Dirichlet prior on π_j, for $j = 1, ..., d$.

under the study. The finite mixture model has the density functions as follows [16]:

$$p(y_i) = \sum_{\ell=1}^{L} \pi_\ell \, p_\ell(y_i), \qquad (1)$$

in which $i = 1, 2, ..., n$, L is the number of mixture components and for each ℓ, $p_\ell(y_i)$ is the density and π_ℓ is the non-negative quantity which amounts to one, that is:

$$\sum_{\ell=1}^{L} \pi_\ell = 1. \qquad (2)$$

The amount of $\pi_1, \ldots, \pi_L$ is called as mixture proportion. $p_1(y_i), \ldots, p_L(y_i)$ is the probability density functions (pdf), thus $p_\ell(y_i)$ is called as mixture density component.

BMM based on model (1), would be applicable when $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$ is a random sample of the D - th dimension of a binary vector. The goal is to divide $\mathbf{Y}$ into L (might be unknown but limited) partition. The L is assumed as the finite mixture density and the BMM can be written as [17]:

$$p(\mathbf{Y} \mid L, \boldsymbol{\pi}, \boldsymbol{\Theta}) = \sum_{\ell=1}^{L} \pi_\ell \, p_\ell(\mathbf{Y} \mid \theta_\ell), \qquad (3)$$

which p_ℓ is called with the mixture density component, $\boldsymbol{\Theta} = (\theta_1, \ldots, \theta_L)$ is a mixture parameter and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_L)$ is the mixture proportion.

For example, supposed that $Y_i, i = 1, 2, ..., n$ is a random variable of binary data and its components are assumed to be independent, where $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iD})$ with $Y_{id} \in [0,1]^D$, and the mixture density component of p_ℓ is Bernoulli distributed with each component is independent. There is $\boldsymbol{\theta}_\ell = (\theta_{\ell 1}, \ldots, \theta_{\ell D})$ for $\ell = 1, \ldots, L$ with $0 \le \theta_{\ell d} \le 1$, $d = 1, \ldots, D$, given $\boldsymbol{\Psi} = (\boldsymbol{\pi}, \boldsymbol{\Theta})$ and when L is unknown, but finite, then $\boldsymbol{\Psi}_\ell = (\boldsymbol{\pi}_\ell, \boldsymbol{\Theta})$. The BMM of vector $\mathbf{Y}_i$, therefore, is independently taken depend on $\boldsymbol{\Psi}_\ell$, and the model can be written as follows [17]:

$$p(\mathbf{Y}_i \mid \boldsymbol{\Psi}_\ell) = \sum_{\ell=1}^{L} \pi_\ell \prod_{d=1}^{D} \theta_{\ell d}^{Y_{id}} (1 - \theta_{\ell d})^{1 - Y_{id}}. \qquad (4)$$

Bernoulli Mixture Regression Model

Bernoulli mixture regression model (BMRM) is developed based on Mixture of Generalized Linear Model which is called Mixture of Generalized Linear Regression Model [18]. In the generalized linear model framework, a random variable Y_i that is named as dependent variable, has a linear relationship with covariates $X_1, X_2, \ldots, X_p$ as follows

$$\eta_i = g(\mu_i) = g(\mathrm{E}(Y_i \mid X)) = \sum_{j=1}^{p} \beta_j X_{ij}, \qquad (5)$$

where η is linear predictor, $g(.)$ is the link function, μ_i is expected value of random variable Y_i and β is regression parameter.

The linear relationship on equation (5) allows the dependent variable distribution is assumed to be the form of the exponential family distributions (i.e. Gaussian, Poisson, Gamma, or Bernoulli). Distinct link functions can be used to perform that relationship. Canonical link function

is one of natural link function which is determined by the exponential of the response's density function. For Bernoulli distribution, the canonical link function is the logit function which can be defined as

$$g(\mu_i) = \text{logit}(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right). \quad (6)$$

Therefore equation (5) can be represented as

$$g(\mu_i) = \text{logit}(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right) = \sum_{j=1}^{p} \beta_j X_{ij}\,, \quad (7)$$

and equation (3) can be redefined as

$$p(\mathbf{Y} \mid L, \boldsymbol{\pi}, \mathbf{X}, \boldsymbol{\beta}) = \sum_{\ell=1}^{L} \pi_\ell \, p_\ell(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\beta}_\ell)\,, \quad (8)$$

Where

$$p_\ell(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\beta}_\ell) \square\ Be(\text{logit}_\ell(\boldsymbol{\mu}))\,, \quad (9)$$

which mean $p_\ell(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\beta}_\ell)$ is Bernoulli distributed with parameter $\text{logit}_\ell(\boldsymbol{\mu})$, $\boldsymbol{\mu} = (\mu_1, \mu_2, ..., \mu_n)'$, $\mathbf{X}' = (\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n)$ and $\boldsymbol{\beta}_\ell = (\beta_1, \beta_2, \ldots, \beta_p)'$.

Bayesian MCMC

Let $\boldsymbol{\Theta} = (\Theta_1, \Theta_2, ..., \Theta_d)' = (\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_L, \boldsymbol{\pi})'$ denote all unknown parameters appearing in the mixture of Bernoulli regression model. By using Bayes theorem, the posterior probability distribution $f(\boldsymbol{\Theta} \mid \mathbf{Y}, L, \mathbf{X})$ can be represented as

$$f(\boldsymbol{\Theta} \mid \mathbf{Y}, L, \mathbf{X}) = \frac{f(\mathbf{Y} \mid \boldsymbol{\Theta}, L, \mathbf{X})\, p(\boldsymbol{\Theta})}{f(\mathbf{Y})} \propto f(\mathbf{Y} \mid \boldsymbol{\Theta}, L, \mathbf{X})\, p(\boldsymbol{\Theta}) \quad (10)$$

where $p(\boldsymbol{\Theta})$ is the prior distribution of $\boldsymbol{\Theta}$ and $f(\mathbf{Y} \mid \boldsymbol{\Theta}, L, \mathbf{X})$ is the mixture likelihood.

In the Bayesian inference approach, parameter estimation processes are performed by integrating the posterior distribution. The integration can be conducted numerically by simulation procedure which is generally recognized as Markov Chain Monte Carlo (MCMC) method.

Generally the Markov Chain Monte Carlo method works with the following steps [19, 20].

1. Select an initial value $\theta^{(0)}$.
2. Generate value $\theta^{(t)}$, $t = 1, 2, \ldots, T$ until the equilibrium condition of the distribution is reached
3. Monitor the convergence of the algorithm using convergence diagnostics. If convergence diagnostics fail for certain *T*, then generate more observations by increasing bigger *T*.
4. Cut off the first *B* observations as a burn-in condition.
5. Consider $\left\{\theta^{(B+1)}, \theta^{(B+2)}, \ldots, \theta^{(T)}\right\}$ as the sample for the posterior analysis.
6. Plot the posterior distribution (usually focus is on the univariate marginal distributions).
7. Finally, obtain summaries of the posterior distributions of θ.

2. Methods

Source of Data

The data used in this research was gathered from Database of Ministry of Research and Technology and Higher Education through Bidikmisi channel, that was Bidikmisi data of all provinces in Indonesia on 2015.

Research Flowchart

The classification analysis procedures done in this paper, i.e. BBMRM, dummy regression model, and polytomous regression model, are given in the following research flows as in Figure 1.

Research Variables

Research variables used in this study consisted of the response variable (*Y*) and the predictor variable (*X*) as follows

Y = the acceptance status of Bidikmisi scholarship (1 = accepted, 0 = not accepted).

X_1 = father's job is formed by dummy variables d_{11}, d_{12}, d_{13} and d_{14}, where

d_{11} : farmer, fisherman or others job which relate with agriculture.
- d_{11} = 1, if father's job is a farmer, fisherman or others job which relate with agriculture.
- d_{11} = 0, otherwise.

d_{12} : civil servants, police and army.
- d_{12} = 1, if father's job is a civil servants, police or army.
- d_{12} = 0, otherwise.

d_{13} : entrepreneur.

d_{13} = 1, if father's job is an entrepreneur.
d_{13} = 0, otherwise.

d_{14} : private employees.
d_{14} = 1, if father's job is a private employees.
d_{14} = 0, otherwise.

X_2 = mother's Job is formed by dummy variables d_{21}, d_{22}, d_{23} and d_{24}, where

d_{21} : farmer, fisher and others job which relate with agriculture.
d_{21} = 1, if mother's job is a farmer, fisher or others job which relate with agriculture.
d_{21} = 0, otherwise.

d_{22} : civil servants, police and army.
d_{22} = 1, if mother's job is a civil servants, police or army.
d_{22} = 0, otherwise.

d_{23} : entrepreneur.
d_{23} = 1, if mother's job is an entrepreneur.
d_{23} = 0, otherwise.

d_{24} : private employees.
d_{24} = 1, if mother's job is a private employees.
d_{24} = 0, otherwise.

X_3 = father's education is formed by dummy variables d_{31}, d_{32} and d_{33}, where

d_{31} : not continue to school.
d_{31} = 1, if father's education is not continue to school.
d_{31} = 0, otherwise.

d_{32} : elementary, junior high or senior high school graduate level.
d_{32} = 1, if father's education is an elementary, junior high or senior high school graduate level.
d_{32} = 0, otherwise.

d_{33} : higher education level.
d_{33} = 1, if father's education is a higher education level.
d_{33} = 0, otherwise.

X_4 = mother's education is formed by dummy variables d_{41}, d_{42} and d_{43}, where

d_{41} : not continue to school.
d_{41} = 1, if mother's education is not continue to school.
d_{41} = 0, otherwise.

d_{42} : elementary, junior high or senior high school graduate level.
d_{42} = 1, if mother's education is an elementary, junior high or senior high school graduate level.
d_{42} = 0, otherwise.

d_{43} : higher education level.
d_{43} = 1, if mother's education is a higher education level.
d_{43} = 0, otherwise.

Three variables in the Bidikmisi enrollment, i.e. "father's income", "mother's income", and "family dependent" are used for forming a Bernoulli mixture distribution as a response of BBMRM. These three variables, therefore, are not used in modeling either on BBMRM, dummy regression model, or polytomous regression model. There are still many variables in the registration form of Bidikmisi, but these four variables selected above are more fundamental variables in considering the acceptance of these grantees, in accordance with one of the rules of acceptance of Bidikmisi is that the income per-capita in the family is no more than certain values.

Research Design: Pre-Processing Stage

The explanations of the techniques used in the pre-processing stages of identification with Bernoulli mixture distribution are as following steps:

Step 1. Taking response variable (Y).

Step 2. Selecting covariate "father's income", "mother's income" and "family dependent".

Step 3. Creating a new covariate by counting the amount of "father's income" and "mother's income" divided by "the number of family dependents", then name it with family income per capita.

Step 4. Coding the covariate family income per capita with the following criteria:
If family income per capita > Rp. 750,000, then the family is categorized as wealthy family which has code of family category, CFC = 0.
If family income per capita < Rp. 750,000 then the family fall into poor family which has code of family category, CFC = 1.

Step 5. Matching the response variable (Y) to the CFC in Step 4 and to the AC (Acceptance Condition) with the Bidikmisi acceptance classification table of "wrong" and "right" which are given on Table 1.

The pre-processing stage result describes response data of the Bernoulli mixture distribution with two mixture components, namely component of wrong acceptance condition and component of right acceptance condition.

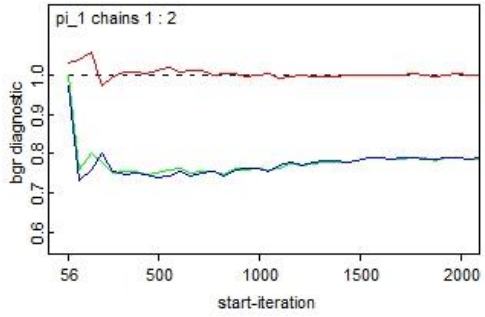

Figure. 4. The convergence of $\hat{\pi}_1$.

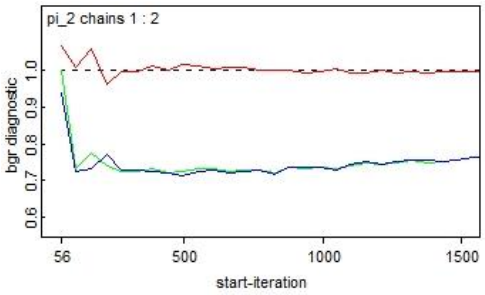

Figure. 5. The convergence of $\hat{\pi}_2$.

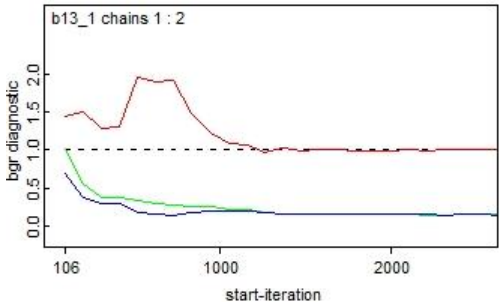

Figure. 6. The convergence of $\hat{\beta}_{13}$ in the first mixture component.

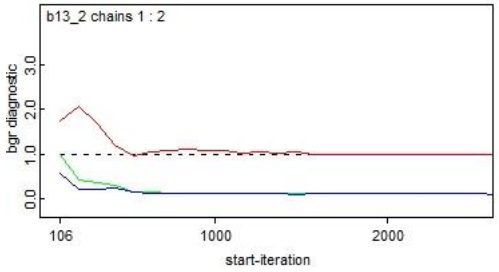

Figure. 7. The convergence of $\hat{\beta}_{13}$ in the second mixture component.

Proposed Model

Referring to equation (8), the two components of BMRM which has to be estimated is defined by

$$f(\mathbf{y}\mid\boldsymbol{\pi},\mathbf{x},\boldsymbol{\beta}) = \pi_1 p_1(y\mid\mathbf{x},\boldsymbol{\beta}_1) + \pi_2 p_2(y\mid\mathbf{x},\boldsymbol{\beta}_2) = \pi_1 Be\left(\frac{e^{g_1(\mathbf{x})}}{1+e^{g_1(\mathbf{x})}}\right) + \pi_2 Be\left(\frac{e^{g_2(\mathbf{x})}}{1+e^{g_2(\mathbf{x})}}\right) \quad (11)$$

with π_1 and π_2 are mixture proportions which have properties $0\le\pi_1\le 1$, $0\le\pi_2\le 1$ and $\pi_1+\pi_2=1$. $f(\mathbf{y}\mid\boldsymbol{\pi},\mathbf{x},\boldsymbol{\beta})$ represents the Bernoulli mixture distribution of two mixture components namely $\pi_1 Be\left(\frac{e^{g_1(\mathbf{x})}}{1+e^{g_1(\mathbf{x})}}\right)$ as component of wrong acceptance condition and $\pi_2 Be\left(\frac{e^{g_2(\mathbf{x})}}{1+e^{g_2(\mathbf{x})}}\right)$ as component of right acceptance condition, where $g_1(\mathbf{x})$ and $g_2(\mathbf{x})$ are determined by

$$\begin{aligned} g_1(\mathbf{x}) &= \beta_{(1)0}+\beta_{(1)11}\mathrm{d}_{11}+\beta_{(1)12}\mathrm{d}_{12}+\beta_{(1)13}\mathrm{d}_{13} \\ &\quad +\beta_{(1)14}\mathrm{d}_{14}+\beta_{(1)21}\mathrm{d}_{21}+\beta_{(1)23}\mathrm{d}_{23}+\beta_{(1)24}\mathrm{d}_{24} \\ &\quad +\beta_{(1)31}\mathrm{d}_{31}+\beta_{(1)32}\mathrm{d}_{32}+\beta_{(1)33}\mathrm{d}_{33} \\ &\quad +\beta_{(1)41}\mathrm{d}_{41}+\beta_{(1)42}\mathrm{d}_{42}+\beta_{(1)43}\mathrm{d}_{43} \end{aligned}$$

and

$$\begin{aligned} g_2(\mathbf{x}) &= \beta_{(2)0}+\beta_{(2)11}\mathrm{d}_{11}+\beta_{(2)12}\mathrm{d}_{12}+\beta_{(2)13}\mathrm{d}_{13} \\ &\quad +\beta_{(2)14}\mathrm{d}_{14}+\beta_{(2)21}\mathrm{d}_{21}+\beta_{(2)23}\mathrm{d}_{23}+\beta_{(2)24}\mathrm{d}_{24} \\ &\quad +\beta_{(2)31}\mathrm{d}_{31}+\beta_{(2)32}\mathrm{d}_{32}+\beta_{(2)33}\mathrm{d}_{33} \\ &\quad +\beta_{(2)41}\mathrm{d}_{41}+\beta_{(2)42}\mathrm{d}_{42}+\beta_{(2)43}\mathrm{d}_{43}. \end{aligned}$$

BMRM Algorithm.

Gibbs sampler is one of algorithms that is frequently used as generator of random variables in MCMC [20]. One advantage of the Gibbs sampler is that, in each step, random values only consider to be generated from univariate conditional distributions. Based on [20], the general Gibbs sampler algorithm for BMRM can be summarized by the following steps on Figure 2.

3. Results and Analysis

In model (11), there are two parameters π_ℓ and $\beta_{\ell p}$ which should be estimated. In order to

TABLE 2
ACCEPTED QUALIFICATION PERCENTAGE

No.	Province	Dummy	Polytomous	BMMRM
1	Aceh	9.16	60.54	71.14
2	North Sumatera	16.69	62.25	82.71
3	West Sumatera	1.24	54.46	98.19
4	Riau	22.17	57.01	73.04
5	Riau Islands	51.11	55.11	63.56
6	Jambi	20.77	56.08	79.82
7	South Sumatera	23.40	57.03	76.49
8	Bangka Belitung	21.98	56.04	85.16
9	Bengkulu	20.77	56.08	79.82
10	Lampung	11.71	65.22	88.11
11	DKI Jakarta	36.62	46.80	74.24
12	West Java	27.39	61.67	84.49
13	Banten	22.90	35.81	96.28
14	Central Java	20.75	62.95	76.85
15	DI Yogyakarta	31.71	19.16	66.03
16	East Java	66.50	47.17	71.70
17	Bali	57.96	52.44	71.50
18	West Nusa Tenggara	7.30	33.15	73.78
19	East Nusa Tenggara	8.75	27.60	92.00
20	West Kalimantan	15.94	47.17	85.16
21	Central Kalimantan	55.5	38.24	77.91
22	South Kalimantan	29.69	45.73	74.33
23	East Kalimantan	24.31	41.39	72.73
24	North Kalimantan	43.75	7.50	86.25
25	North Sulawesi	63.34	44.57	91.64
26	West Sulawesi	60.74	26.16	85.05
27	Central Sulawesi	22.34	45.39	84.21
28	Southeast Sulawesi	51.19	39.53	67.46
29	South Sulawesi	6.03	30.43	77.32
30	Gorontalo	67.46	49.15	97.18
31	Maluku	23.02	57.24	84.21
32	North Maluku	51.69	40.58	61.84
33	West Papua	32.07	47.17	60.38
34	Papua	27.81	36.77	61.56

make inference in Bayesian perspective, the prior distributions for each parameter is $p(\pi_\ell) \sim Dirichlet(\vartheta)$ or π_ℓ has Dirichlet distribution with parameter ϑ, and $p(\beta_{\ell p}) \sim N(\mu,\sigma)$ or $\beta_{\ell p}$ has Normal distribution with parameter μ and σ. Those prior distributions are used to determine the posterior model of $f\left(\mathbf{y}\,|\,\boldsymbol{\pi},\mathbf{x},\boldsymbol{\beta}\right)$. In this research, we define BMRM which is estimated by Bayesian approach as BBMRM.

BBMRM Algorithm.

The Gibbs sampler algorithm for BBMRM can be constructed by the following steps on Figure 3. This algorithm could be performed on OpenBUGS software [21] to estimate BBMRM for each province.

For example, BBMRM for the province of East Java has a significant estimated model (12).

$$\hat{f}\left(\mathbf{y}\,|\,\boldsymbol{\pi},\mathbf{x},\boldsymbol{\beta}\right) = 0.6041\; Be\left(\left[\frac{e^{\hat{g}_1(\mathbf{x})}}{1+e^{\hat{g}_1(\mathbf{x})}}\right]\right) + 0.3959\; Be\left(\left[\frac{e^{\hat{g}_2(\mathbf{x})}}{1+e^{\hat{g}_2(\mathbf{x})}}\right]\right) \quad (12)$$

$$\begin{aligned}\hat{g}_1(x) = {} & 1.19 - 1.287d_{11} - 1.204d_{12} - 1.094d_{13} \\ & - 0.77d_{14} - 1.99d_{21} - 1.50d_{23} - 1.29d_{24} \\ & - 0.69d_{31} - 0.218d_{32} + 0.19d_{33} \\ & - 0.256d_{41} - 0.116d_{42} + 0.08d_{43}\end{aligned}$$

and

$$\begin{aligned}\hat{g}_2(x) = {} & -1.72 + 0.87d_{11} - 0.066d_{12} + 0.611d_{13} \\ & + 0.04d_{14} + 1.13d_{21} + 0.58d_{23} + 0.067d_{24} \\ & + 0.086d_{31} + 0.102d_{32} + 0.053d_{33} \\ & - 0.069d_{41} - 0.15d_{42} + 0.005d_{43}\end{aligned}$$

In order to have the valid posterior inference for parameters, the Markov chains of estimated parameters should be convergent which implies that the chains reaches the posterior distribution. The MCMC convergence of estimated parameter can be monitored through Brooks-Gelman-Rubin method [21].

In relation with estimation processes of model (12), Figure 4 presents the convergence of $\hat{\pi}_1$, whereas Figure 5 shows the convergence of $\hat{\pi}_2$. These graphics describe the evolution of the pooled posterior variance which has green color line, average within-sample variance which is plotted as a blue color line, and their ratio which is marked in red line. The ratio which converge to one means that the estimated parameter is

convergent to posterior distribution.

Similar convergence results are achieved for estimated parameters $\beta_{\ell p}$. For example convergence of $\hat{\beta}_{13}$ in first mixture component as shown on Figure 6 and convergence of $\hat{\beta}_{13}$ in second mixture component as shown on Figure 7.

The calculation results of accepted qualification percentage for all province are presented on Table 2. Those three models are built by using 70 percent of the data as the in-sample or training data. While, the other 30 percent of data, which is set to have a randomly representative member, as an out-sample for model validation. By regarding to Table 2, it can be shown that the accuracy of classification on dummy regression model, polytomous regression model, and BBMRM are about between 1% - 67%, 7% - 65%, and 60% - 98% respectively. There are evidence that BBMRM is more accurate to clasify in 31 provinces than polytomous regression model and on 27 provinces than dummy regression model.

4. Conclusion

BBMRM couple with MCMC approach gives higher percentage of acceptance classification accuracy than dummy regression model and polytomous regression model. This BBMRM is more representative for Bidikmisi acceptance modeling. As a future research, three methods for Bidikmisi classification discussed in this paper can be compared with other existing classification method, i.e. the Classification and Regression Trees (CART) approach, Neural Network, and Support Vector Machine (SVM). The validation of the three methods obove is still done on 30 percent of the out-sample data in the same year of Bidikmisi selection. In the next study, validation of BBMRM can be done for Bidikmisi data in the following year. Further research can also be done on the existence of the influence of covariate on the component weight, π_j, of BBMRM.

Acknowledgements

The Authors are grateful to Directorate for Research and Community Service (DRPM) Ministry of Research, Technology and Higher Education Indonesia which support this research under PUPT research grant no 608/PKS/ITS/2017.

References

[1] Anon, Pedoman Penyelenggaraan Bantuan Biaya Pendidikan Bidikmisi Tahun 2016, Jakarta: Direktorat Jenderal Pembelajaran dan Kemahasiswaan, Kementerian Riset Teknologi dan Pendidikan Tinggi, 2016.

[2] X. Wang and A. Kabán, "Finding Uninformative Features in Binary Data," *Intelligent Data Engineering and Automated Learning - IDEAL 2005,* vol. 3578, pp. 40–47, 2005.

[3] R. Duda and P. Hart, Pattern Classification and Scene Analysis, Wiley, 1973.

[4] J. Grim, P. Pudil and P. Somol, "Multivariate Structural Bernoulli Mixtures for Recognition of Hand written Numerals," in *Proceedings 15th International Conference on Pattern Recognition (ICPR-2000), vol. 4, pp. 585–589.*, Barcelona, Spain, Spain, 2000.

[5] J. González, A. Juan, P. Dupont, E. Vidal and F. Casacuberta, "A Bernoulli Mixture Model for Word Categorization," in *Proceedings of the IX Spanish Symposium on Pattern Recognition and Image Analysis*, Benicassim, Spain, 2001.

[6] A. Juan and E. Vidal, "On The Use of Bernoulli Mixture Models for Text Classification," *Pattern Recognition,* vol. 35, no. 12, pp. 2705–2710, 2002.

[7] A. Juan and E. Vidal, "Bernoulli Mixture Models for Binary Images," in *Proceeding of the 17th International Conference on Patern Recognition (ICPR'04)*, 2004.

[8] A. Patrikainen and H. Mannila, "Sub Space Clustering of High-Dimensional Binary Data-A Probabilistic Approach," Workshop on Clustering High-Dimensional Data and Its Applications, 2004.

[9] S. Zhu, I. Takigawa, S. Zhang and H. Mamitsuka, "A Probabilistic Model for Clustering Text Documents with Multiple Fields," in *Advances in Information Retrieval, 29th European Conference on IR Research (ECIR2007)*, Berlin, Heidelberg, 2007.

[10] Z. Sun, O. Rosen and A. Sampson, "Multivariate Bernoulli Mixture Models with application to Postmortem Tissue Studies in Schizophrenia," *Biometrics,* vol. 63, pp. 901-909, 2007.

[11] J. Tikka, J. Hollmen and S. Myllykangas, "Mixture Modelling of DNA Copy Number Amplification Patterns in Cancer," in *Proceedings of the 9th International Work-Conference on Artificial Neural Networks (IWANN'2007)*, Springer-Verlag, Berlin,

Heidelberg, 2007.

[12] S. Myllykangas, J. Tikka, T. Böhling, S. Knuutila and J. Hollmén, "Classification of Human Cancers Based on DNA Copy Number Amplification Modelling," *BMC Med. Genomics,* vol. 1, pp. 1-13, 2008.

[13] J. Hollmen and J. Tikka, "Compact and Understandable Descriptions of Mixture of Bernoulli Distributions," in *Proceedings of 7th International Symposium on Intelligent Data Analysis (IDA2007)* , Springer-Verlag, Berlin, Heidelberg, 2007.

[14] M. Saeed, K. Javed and H. A. Babri, "Machine Learning Using Bernoulli Mixture Models: Clustering,Rule Extraction and Dimensionality Reduction,"*Neuro-computing,* vol. 119, pp. 366–374, 2013.

[15] N. Bouguila, "On Multivariate Binary Data Clustering and Feature Weighting," *Comput. Stat.Data Anal,* vol. 54, pp. 120-134, 2010.

[16] A. B. Astuti, N. Iriawan, Irhamah, and H. Kuswanto, "Bayesian Mixture Model Averaging for Identifying the Different Gene Expressions of Chickpea (Cicer Arietinum) Plant Tissue", *Communications in Statistics - Theory and Methods*, Vol. 46, Issue 21, pp 10564-10581,**DOI:** 10.1080/03610926.2016. 1239112, 2017.

[17] M. Nadif and G. Govaert, "Clustering for Binary Data and Mixture Models-Choice of The Model," *Applied Stochastic Models in Business and Industry,* vol. 13, no. 3-4, pp. 269-278, 1997.

[18] B. Grun and F. Leisch, "Finite mixtures of generalized linear regression models," in *Recent Advances in Linear Models and Related Areas*, Shalabh and C. Heumann, Eds., Heidelberg, Springer, 2008, pp. 205 - 230.

[19] I. Ntzoufras, Bayesian modeling using WinBUGS, New york. : John Wiley & Sons, 2009.

[20] G. Casella and E. George, "Explaining Gibss Sampler," *The America Statistical Association,* vol. 46, no. 3, pp. 167-174, 1992.

[21] D. Lunn, D. Spiegelhalter, A. Thomas and N. Best, "The BUGS project: evolution, critique and future directions (with discussion).," *Statistics in Medicine,* vol. 28, pp. 3049 - 3082. DOI : 10.1002/sim.3680, 2009.